ON SOME OPTIMAL SAMPLING PROCEDURES FOR SELECTION PROBLEMS

Shanti S. Gupta, et al

Purdue University

Prepared for:

Office of Naval Research

September 1974

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Mimeograph Series #394	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: AD/A-000 255	
4. TITLE (and Subtitle): On some Optimal Sampling Procedures for Selection Problems	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER: Mimeograph Series #394
7. AUTHOR(s): Shanti S. Gupta and Deng-Yuan Huang	8. CONTRACT OR GRANT NUMBER(s): N00014-67-A-0226-00014
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Purdue University, West Lafayette, Indiana 47907	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 00014
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Washington, DC	12. REPORT DATE: September, 1974 13. NUMBER OF PAGES: 13
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public,release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

optimal sampling
selection procedures
minimax
Bayes

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper deals with determining the minimum sample sizes for selecting the population with the largest location (or scale) parameter based on the Γ-minimax criterion.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

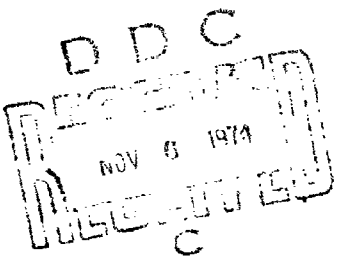
On Some Optimal Sampling Procedures for Selection Problems*

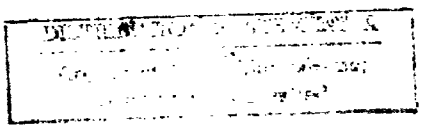
by

Shanti S. Gupta and Deng-Yuan Huang
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #394

September 1974





*This research was supported by the Office of Naval Research Contract N00014-67-A-0226-00014 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

On Some Optimal Sampling Procedures for Selection Problems*

by

Shanti S. Gupta and Deng-Yuan Huang
Purdue University

1. Introduction

An experimenter is asked which of k populations has the largest mean and must decide how large a sample he should take to decide this question. Taking a large sample decreases the probability of an incorrect decision, but at the same time increases the cost of sampling. It seems reasonable that the "optimum" sample size should depend both on the cost of sampling and the amount of use to be made of the decision. In this paper, loss functions are set up which take into consideration the amount of use to be made of the result, the cost of making a wrong decision and the cost of sampling. Assume that the parameters are random variables and only partial prior information is available. The Γ-minimax criterion allows one to determine a sample size that minimizes the maximum expected risk over Γ which is a class of prior distributions. Note that if Γ consists of a single prior, then the Γ-minimax criterion is the Bayes criterion for that prior. At the other extreme, if Γ consists of all priors then the Γ-minimax criterion is the usual minimax criterion. Some statements for the development of Γ-minimax criterion have been discussed by Gupta and Huang [3]. Dunnett [1] discussed an optimal sampling problem for the

---

*This research was supported by the office of Naval Research Contract N00014-67-A-0226-00014 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

normal means problem by the usual minimax and Bayes criterion. Ofosu [4] considered the minimax criterion for gamma populations. In Section 2, we discuss the location parameter problem. Sclae parameter problem is considered in Section 3.

2. Location Parameters

Let there be $k(\geq 2)$ independent populations with continuous distribution functions $F(x-\theta_1)$, $F(x-\theta_2),\ldots,F(x-\theta_k)$, where the location parameters θ_i's are unknown. Let $X_{i1},\ldots,X_{in}$ denote n independent observations from the ith population and define $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$, $(1 \leq i \leq k)$. It is well known that the distribution of $X_{ij} - \theta_i$ does not depend on $\theta_i (1 \leq i \leq k)$. Define

$$G_n(y|F) = P\{\frac{1}{n}[(X_{i1}-\theta_i)+\ldots+(X_{in}-\theta_i)] \leq y\}.$$

Then for any i, $1 \leq i \leq k$,

$$P\{\bar{X}_i \leq x\} = G_n(x-\theta_i|F).$$

We wish to select the population associated with the largest θ_i's using the usual selection procedure. We wish to determine the optimal sampling by Γ-minimax criterion. Let δ_i denote the probability of selecting the ith population. Define the usual procedure as follows:

$$\delta_i(\underline{x}) = \begin{cases} 1 & \text{if} \quad \bar{x}_i \geq \max_{\substack{1\leq j\leq k \\ j\neq i}} \bar{x}_j, \\ 0 & < \end{cases}.$$

Then the probability p_i that the ith population is selected is given by

$$P_i = P\{\max_{\substack{1\le j\le k\\ j\neq i}} \bar{X}_j \le \bar{X}_i\}$$

$$= \int_{-\infty}^{\infty} \sum_{\substack{j=1\\ j\neq i}}^{k} G_n(x+\theta_i-\theta_j|F)\, dG_n(x|F).$$

Let $\Omega = \{\underline{\theta}|\underline{\theta} = (\theta_1,\theta_2,\ldots,\theta_k)\}$ and $\Omega_i = \{\underline{\theta}|\theta_i \ge \max_{\substack{1\le j\le k\\ j\neq i}} \theta_j + \Delta\}$, $i=1,2,\ldots,k$, and Δ is a given positive constant. Then $\Omega = \Omega_0 \cup \Omega_1 \cup \ldots \cup \Omega_k$, where the Ω_0 is that part of Ω usually called indifference zone.

For $\underline{\theta} \in \Omega_i$, $1 \le i \le k$, define $L^{(r)}(\underline{\theta},\delta_j) = 0$ for all $r\neq i$, $L^{(i)}(\underline{\theta},\delta_j) = c'(\theta_i - \theta_j)\delta_j$, $j = 1,2,\ldots,k$, where $L^{(i)}(\underline{\theta},\delta_j)$ represents the loss for $\underline{\theta} \in \Omega_i$ when the jth population is selected, c' being a positive constant. For $\underline{\theta} \in \Omega_0$, the loss is zero.

The probability of making a wrong decision can be decreased by increasing the size of the experiment on which the decision is to be based, but this increases the cost of experimentation, which must also be considered. It will be assumed here that the cost of performing an experiment involving n observations from each population is cn, where c is a positive constant. Let ρ be a distribution over Ω. Then the risk function, or the expected loss, with experimentation costs included is

$$\gamma_n(\rho) = cn + c' \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{\Omega_i} \int_{E^k} (\theta_i-\theta_j)\delta_j(\underline{x})dF_{\underline{\theta}}(x)d\rho(\underline{\theta}).$$

Assume that partial information is available in the selection problem, so that we are able to specify $\pi_i = P\{\underline{\theta} \in \Omega_i\}$, $\sum_{i=0}^{k} \pi_i = 1$. Define

$$\Gamma = \{\rho(\underline{\theta})|\int_{\Omega_i} d\rho(\underline{\theta}) = \pi_i,\ i = 1,2,\ldots,k\}.$$

We know that

$$\gamma_n(\rho) = cn + c' \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{\Omega_i} (\theta_i - \theta_j) p_j \, d\rho(\underline{\theta}).$$

For any i, $1 \leq i \leq k$, let $\underline{\theta}_i^*$ be some point in Ω_i such that

$$\sup_{\underline{\theta} \in \Omega_i} \sum_{j=1}^{k} (\theta_i - \theta_j) p_j = \sum_{j=1}^{k} (\theta_i^* - \theta_j^*) p_j^*,$$

where $\underline{\theta}_i^* = (\theta_1^*, \theta_2^*, \ldots, \theta_k^*)$ and

$$p_j^* = \int_{-\infty}^{\infty} \prod_{\substack{j=1 \\ j \neq i}}^{k} G_n(x + \theta_i^* - \theta_j^*) \, dG_n(x).$$

Then

$$\sup_{\rho \in \Gamma} \gamma_n(\rho) = cn + c' \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} (\theta_i^* - \theta_j^*) p_j^*.$$

For $\underline{\theta} \in \Omega_i$, $1 \leq i \leq k$, let

$$R^{(i)}(\theta_1, \theta_2, \ldots, \theta_k) = \sum_{j=1}^{k} (\theta_i - \theta_j) p_j,$$

and $g_{ij} = \theta_i - \theta_j$, then

$$R^{(i)}(g_{i1}, g_{i2}, \ldots, g_{ik}) = \sum_{j=1}^{k} g_{ij} \int_{-\infty}^{\infty} \prod_{\substack{j=1 \\ j \neq i}}^{k} G_n(x + g_{ij}) \, dG_n(x).$$

Note that all the g_{ij} are positive, by definition, and $g_{ii} = 0$.

We first require to determine the values of the parameters in Ω_i for which $R^{(i)}$ is a maximum. Somerville [5] shows that this is achieved when $g_{ij} (j \neq i)$ are positive and equal, while $g_{ii} = 0$. Denote the common value of the positive g_{ij} by g_i, then

$$R^{(i)} = \sum_{\substack{j=1 \\ j \neq i}}^{k} g_i p_j = g_i (1 - p_i).$$

If we denote by $R_M^{(i)}$ the maximum, with respect to y, of the function $R^{(i)}$, then the maximum risk is given by

$$\sup_{\rho\in\Gamma} \gamma_n(\rho) = cn + c' \sum_{i=1}^{k} \pi_i R_M^{(i)}.$$

Since

$$R^{(i)} = g_i\{1 - \int_{-\infty}^{\infty} G_n^{k-1}(x + g_i)dG_n(x)\},$$

hence we know that

$$\sup_{g_1\geq\Delta} R^{(1)} = \ldots = \sup_{g_k\geq\Delta} R^{(k)} = \sup_{g\geq\Delta} R = R_M.$$

Thus, $$\sup_{\rho\in\Gamma} \gamma_n(\rho) = cn + c'(\sum_{i=1}^{k} \pi_i)R_M$$

$$= cn + c'(1-\pi_0)R_M.$$

For the problem of normal distributions $N(\theta_i,1)$, $i = 1,2,\ldots,k$, so that we are interested in the selection of the means θ_i, the function R becomes

$$R = g[1-\Phi_{k-1,1/2}(g\sqrt{\tfrac{n}{2}},\ldots,g\sqrt{\tfrac{n}{2}})],$$

where $\Phi_{k-1,1/2}(\cdot)$ is the (k-1)-variate normal distribution with all correlation coefficients equal to $\frac{1}{2}$. If we denote by M_{k-1} the maximum, with respect to y ($\geq \Delta$), of the function

$$y[1-\Phi_{k-1,1/2}(y,\ldots,y)],$$

Then the maximum risk is given by

$$\sup_{\rho\in\Gamma} \gamma_n(\rho) = cn + \frac{1}{\sqrt{\frac{n}{2}}} c'(1-\pi_0)M_{k-1}.$$

This can be minimized with respect to n by taking $n = <(\frac{1}{2c^2} c'^2(1-\pi_0)^2 M_{k-1}^2)^{\frac{1}{3}}>$, where $<x>$ is the smallest integer greater than or equal to x. The values of M_{k-1} have been discussed by Dunnett [1]. The values of $\Phi_{k-1,1/2}(y,\ldots,y)$ are also tabulated in Gupta [2] for k = 2(1)13.

By using Somerville's table [5] for the values of M_{k-1}, k = 2(1)6, we compute some n values as follows:

Table I

Minimum Sample Sizes for the Normal Means Problem

k	π_0 \ $\frac{c'}{c}$	5	10	15	30	50	100
2	0.10	1	2	2	3	4	5
	0.30	1	1	2	2	3	5
	0.50	1	1	1	2	3	4
3	0.10	1	2	2	3	5	7
	0.30	1	2	2	3	4	6
	0.50	1	1	2	2	3	5
4	0.10	2	2	3	4	5	8
	0.30	1	2	2	3	5	7
	0.50	1	2	2	3	4	6
5	0.10	2	2	3	4	6	9
	0.30	1	2	2	4	5	8
	0.50	1	2	2	3	4	6
6	0.10	2	2	3	4	6	9
	0.30	2	2	3	4	5	8
	0.50	1	2	2	3	4	6

Note that we choose $\Delta \leq 0.5$.

3. Scale Parameters

We assume that the k independent populations have continuous distribution functions $F(\frac{x}{\theta_1})$, $F(\frac{x}{\theta_2}),\ldots,F(\frac{x}{\theta_k})$, respectively, where the scale parameter θ_i is positive and unknown and $x \geq 0$. Let $X_{i1},\ldots,X_{in}$ denote n independent observations from π_i and devine $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$, $1 \leq i \leq k$. As before, we know that for any i, $1 \leq i \leq k$,

$$P\{\bar{X}_i \leq x\} = H_n(\frac{x}{\theta_i}|F).$$

Let $\delta_i(\underline{x})$ denote the same procedure as before. Then the probability q_i that π_i is selected is given by

$$\begin{aligned} q_i &= P\{\max_{\substack{1\leq j\leq k\\ j\neq i}} \bar{X}_j \leq \bar{X}_i\} \\ &= \int_{-\infty}^{\infty} \prod_{\substack{j=1\\ j\neq i}}^{k} H_n(x\cdot\frac{\theta_i}{\theta_j}|F)dH_n(x|F). \end{aligned}$$

Let $\Omega_i = \{\underline{\theta}|\theta_i \geq \delta \max_{\substack{1\leq j\leq k\\ j\neq i}} \theta_j\}$, $i = 1,2,\ldots,k$, and $\delta(>1)$ is a given constant.

For $\underline{\theta} \in \Omega_i$, $1 \leq i \leq k$, define $L^{(r)}(\underline{\theta},\delta_j) = 0$ for all $r\neq i$,

$$L^{(i)}(\underline{\theta},\delta_j) = c'\delta_j \log\frac{\theta_i}{\theta_j},\ j = 1,2,\ldots,k,$$

where $L^{(i)}(\underline{\theta},\delta_j)$ represents the loss for $\underline{\theta} \in \Omega_i$ but the jth population is selected, c' being a positive constant. For $\underline{\theta} \in \Omega_0$, the loss is zero.

By using similar discussion as before, for any i, $1 \leq i \leq k$, let $\underline{\theta}_i^*$ be some point in Ω_i such that

$$\sup_{\underline{\theta}\in\Omega_i} \sum_{j=1}^{k} p_j \log\frac{\theta_i}{\theta_j} = \sum_{j=1}^{k} p_j^* \log\frac{\theta_i^*}{\theta_j^*},$$

where $\underline{\theta}_i^* = (\theta_1^*, \theta_2^*, \ldots, \theta_k^*)$ and

$$q_j^* = \int_{-\infty}^{\infty} \prod_{\substack{j=1 \\ j \neq i}}^{k} H_n(x \cdot \frac{\theta_i^*}{\theta_j^*}) dH_n(x).$$

Then

$$\sup_{\rho \in \Gamma} \gamma_n(\rho) = cn + c' \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} q_j^* \log \frac{\theta_i^*}{\theta_j^*},$$

where Γ is defined as before.

Let

$$Q^{(i)}(\theta_1, \theta_2, \ldots, \theta_k) = \sum_{j=1}^{k} q_j \log \frac{\theta_i}{\theta_j},$$

and $h_{ij} = \theta_i/\theta_j$, then

$$Q^{(i)}(h_{i1}, \ldots, h_{ik}) = \sum_{j=1}^{k} (\log h_{ij}) \int_{-\infty}^{\infty} \prod_{\substack{j=1 \\ j \neq i}}^{k} H_n(x+h_{ij}) dG_n(x).$$

Note that all the h_{ij} are positive, by definition, and $h_{ii} = 1$. By using the similar discussion as in Section 2, we have

$$\sup_{\rho \in \Gamma} \gamma_n(\rho) = cn + c'(1-\pi_0) Q_M,$$

where $Q_M = \sup_{h \geq \delta} Q = \sup_{h_1 \geq \delta} Q^{(1)} = \ldots = \sup_{h_k \geq \delta} Q^{(k)} = (\log h)\{1 - \int_0^{\infty} H_n^{k-1}(xh) dH_n(x)\}$.

For the selection of the scale parameters of gamma populations with densities

$$\frac{1}{\theta_i} \left(\frac{x}{\theta_i}\right)^{a-1} \frac{1}{\Gamma(a)} \exp\{-\frac{x}{\theta_i}\}, \quad x > 0,$$

where θ_i, $1 \leq i \leq k$, are the unknown scale parameters, $\theta_i > 0$ and $a(>0)$ is a known shape parameter. Let

$$Q_n = cn + c'(1-\pi_0)(\log h)[1 \int_0^{\infty} G_\nu^{k-1}(h \cdot) dG_\nu(x)],$$

where $G_\nu(x)$ is the cdf of a standardized gamma random variable with $\nu=na$. Let N_{k-1} be the maximum, with respect to $h(\geq\delta)$, of the function

$$[1 - \int_0^\infty G_\nu^{k-1}(hx)dG_\nu(x)]\log h,$$

then the maximum risk is given by

$$Q_M = cn + c'(1-\pi_0)N_{k-1}.$$

It is not analytically feasible to minimize Q_M with respect to n by differentiation. We can use the same method as Ofosu [4] to make a numerical study. Some asymptotic Γ-minimax solutions are discussed as follows.

It is known that as ν tends to ∞,

$$(\frac{2\nu-1}{2})^{\frac{1}{2}} \log\{\bar{X}_i/(a\theta_i)\}, \quad (i = 1,2,\ldots,k)$$

is asymptotically distributed as $N(0,1)$. Hence, as $\nu \to \infty$, we have

$$\sup_{\rho\in\Gamma} \gamma_n(\rho) = cn + c'(1-\pi_0)\, 2(2\nu-1)^{-1}M_{k-1},$$

where $M_{k-1} = \sup_{h\geq\delta} h\{1-\Phi_{k-1,1/2}(h,\ldots,h)\}$. Then $\sup_{\rho\in\Gamma} \gamma_n(\rho)$ can be minimized with respect to n, for large ν, by taking

$$n = <\frac{1}{2}\{\sqrt{\frac{c'}{c}}(1-\pi_0)4M_{k-1}+1\}>,$$

where $<x>$ is the smallest integer greater than or equall to x.

Applications to selection of Weibull populations scale parameters and normal variances problems can be obtained in the same way as in Ofosu [4].

References

[1] Dunnett, C. W. (1960). On selecting the largest of k normal population means. J. Roy. Statist. Soc. Ser. B, 22, 1-40.

[2] Gupta. S. S. (1963). Probability integrals of multivariate normal and multivariate t. Ann. Math. Statist. 34, 792-828.

[3] Gupta S. S. and Huang, D. Y. (1974). On some Γ-minimax subset selection and multiple comparison procedures. Department of Statistics, Purdue University, Mimeo Series #392, W. Laf., IN.

[4] Ofosu, J. B. (1974). A minimax procedure for selecting the population with the largest (smallest) scale parameter. Calcutta Statist. Assoc. Bulletin 16, 143-154.

[5] Somerville, D. N. (1954). Some problems of optimum sampling. Biometrika 41, 420-429.